CARVANA

Don’t Get Kicked

Predict whether Car Purchase at Auction is a ‘Lemon’

BIOS 635: Final Project Report
Junead Khan

What is a ‘Lemon Car’

+

=

CARVANA

Aim:

Can we predict whether a car bought at an auction is a Lemon Car?

CARVANA

The Dataset

Training Dataset:
72,893
records

Test Dataset:
47,707
records

32 Unique Features

PurchDate, Auction, VehYear, VehicleAge, Make, Model, Trim, SubModel, Color, Transmission, WheelTypeID, WheelType, VehOdo, Nationality, Size, TopThreeAmericanName, MMRAcquisitionAveragePrice, MMRAcquisitionCleanPrice, MMRAcquisitionRetailAveragePrice, MMRAcquisitionRetailCleanPrice, MMRCurrentAuctionAveragePrice, MMRCurrentAuctionCleanPrice, MMRCurrentRetailAveragePrice, MMRCurrentRetailCleanPrice, PRIMEUNIT, AcquisitionType, AUCGUART, KickDate, BYRNO, VNZIP, VNST, VehBCost, IsOnlineSale, WarrantyCost

Holding Out Data

Training Dataset:
72,893
records

Split Training Dataset:
2/3

Hold Out Dataset:
1/3

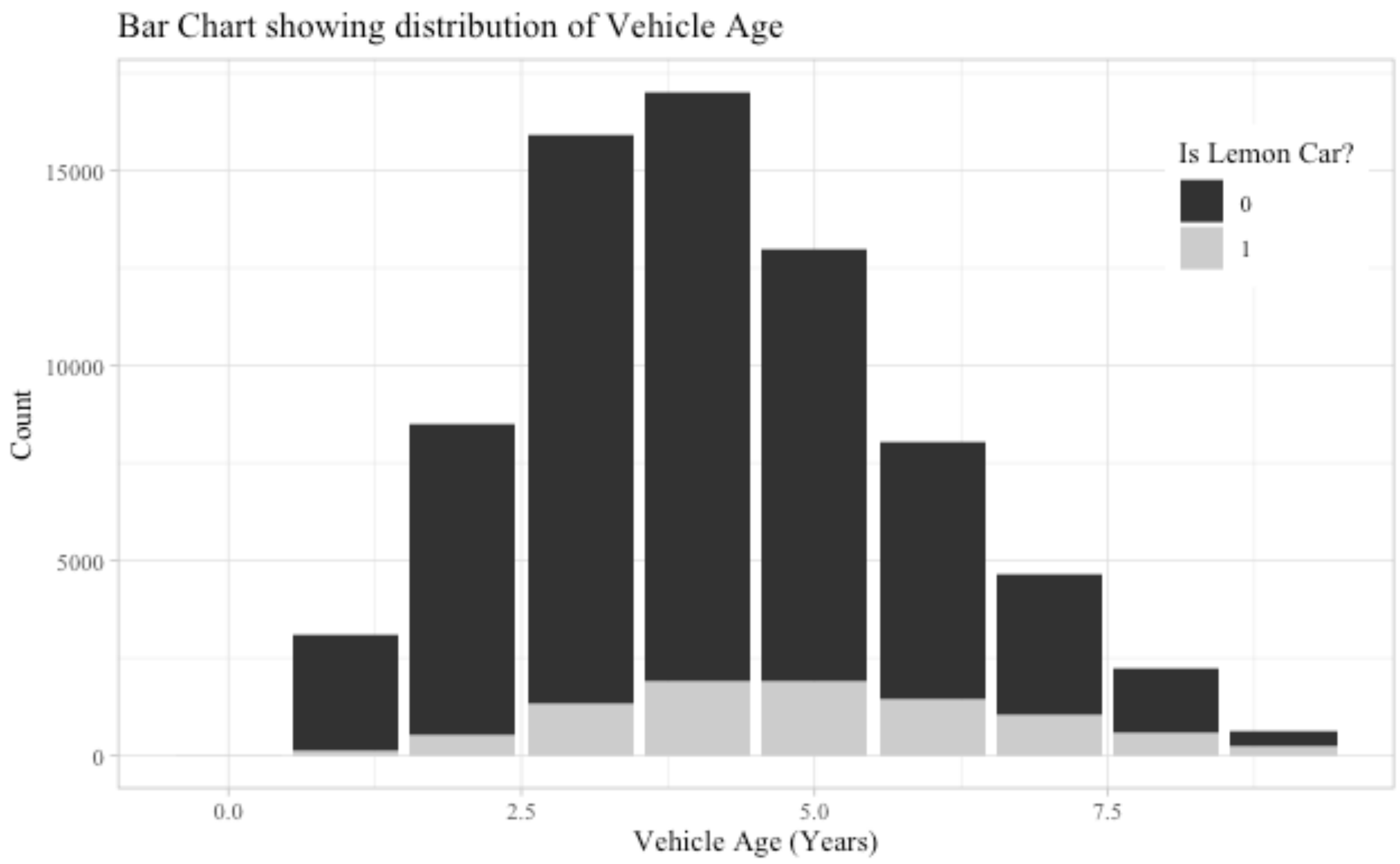
1. Data Cleaning

- Variables with high number of NULL removed (>90%)
- Variables deemed to be not relevant were removed
- Transmission Variable NULL WheelType = ’Steel’ Wheels
- One-to-one mapping between WheelType and WheelTypeID
- Changed String ‘NULL’s to NA in R
- Converted MMR variables to numeric
- Converted other character variables to Factors.

CARVANA

2. Data Exploration

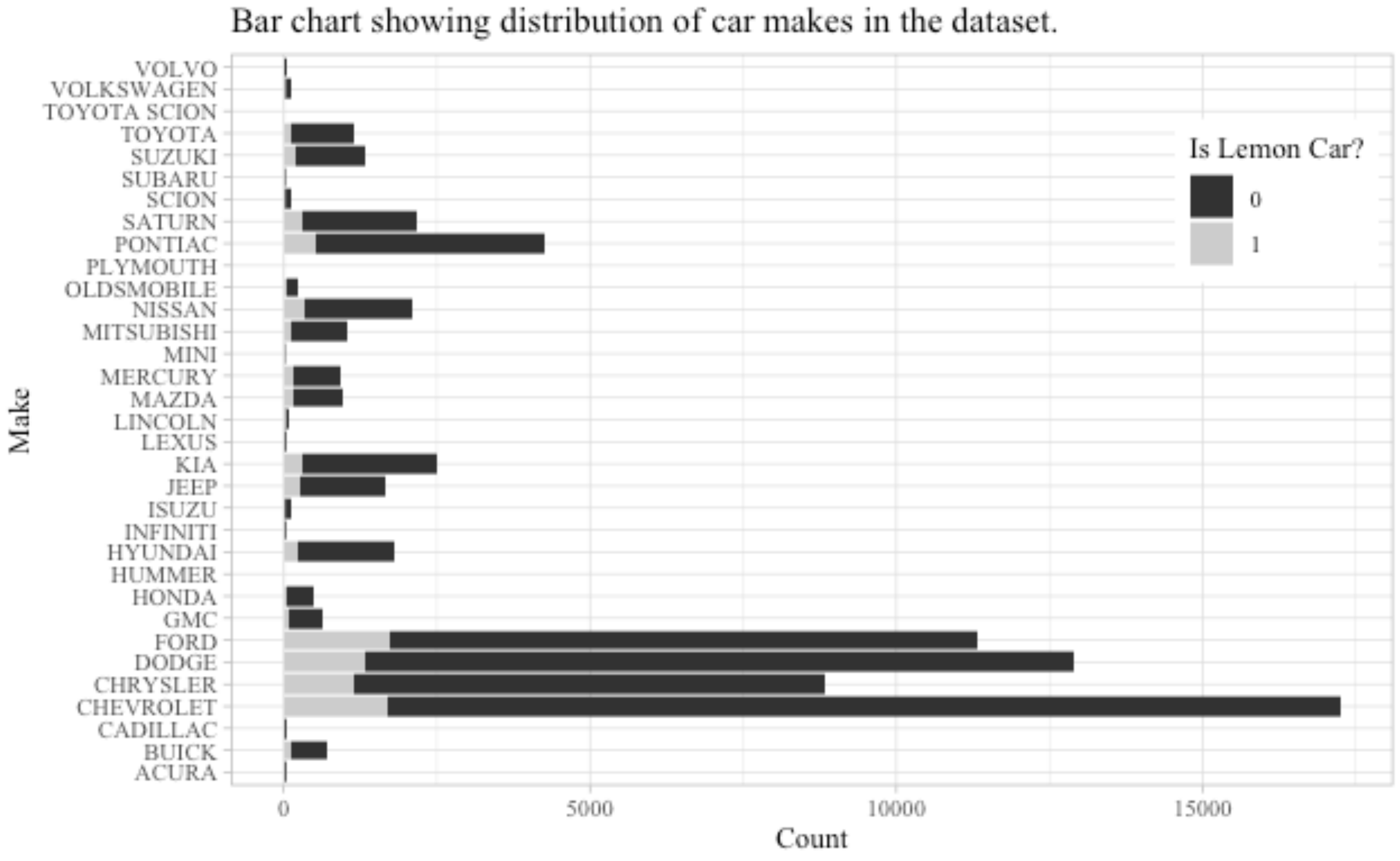
2.1 Vehicle Age



Mean Age: 4.18
Median Age: 4.00

Newest Car: 0 Years
Oldest Car: 9 Years

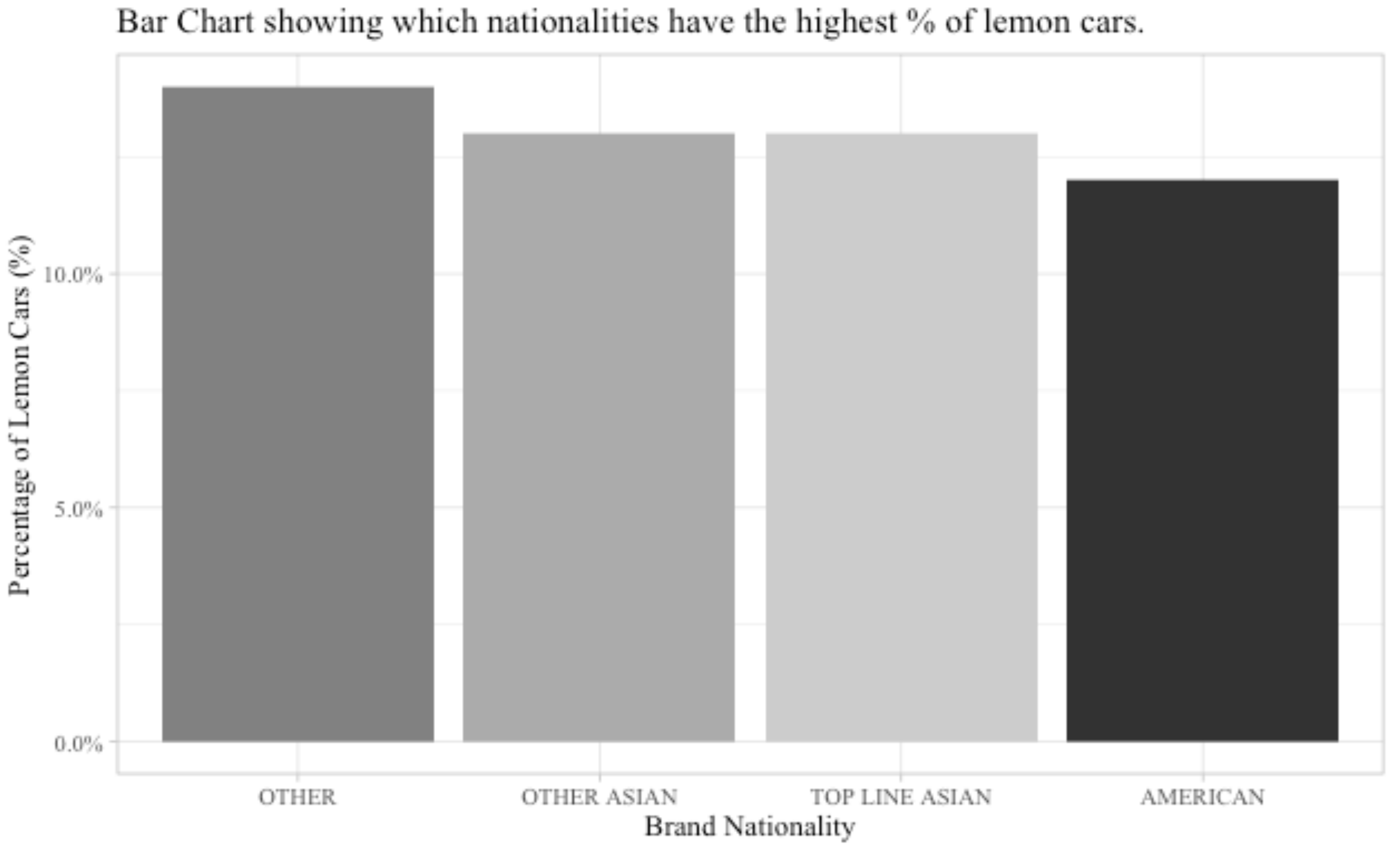
2.2 Vehicle Make



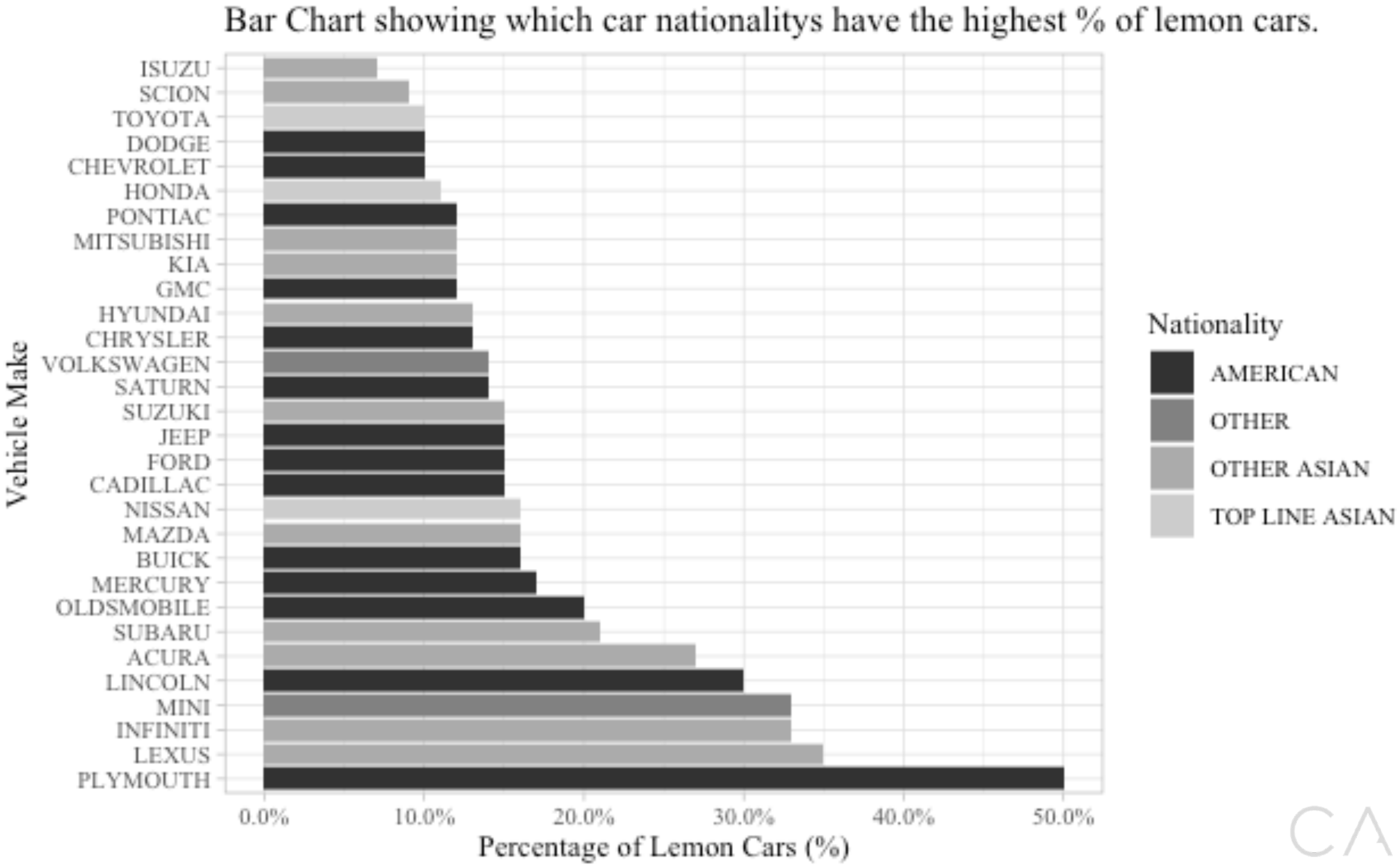
American Manufacturers overrepresented
e.g. Chevrolet, Ford, Chrysler, Dodge, Cadillac

Asian Manufacturers underrepresented
e.g. Toyota, Honda, Nissan, Subaru

2.3 Vehicle Nationality



2.4 Vehicle Make Relative to Proportion



CARVANA

3. Feature Engineering

CARVANA

Using Model to create 2 new features

Model

RAV4 V6 4WD

via RegEx Matching

Displacement V6

Powertrain 4WD

Price Difference

MMRAcquisitionAveragePrice

$12,000

Model

$7,000

Difference →

price_difference

$5000

CARVANA

Miles Travelled Per Year

VehOdo	Vehicle Age		miles_per_year
34,000	5	$\xrightarrow{\text{Ratio}}$	6,800
36,000	2	$\xrightarrow{\text{Ratio}}$	18,000

CARVANA

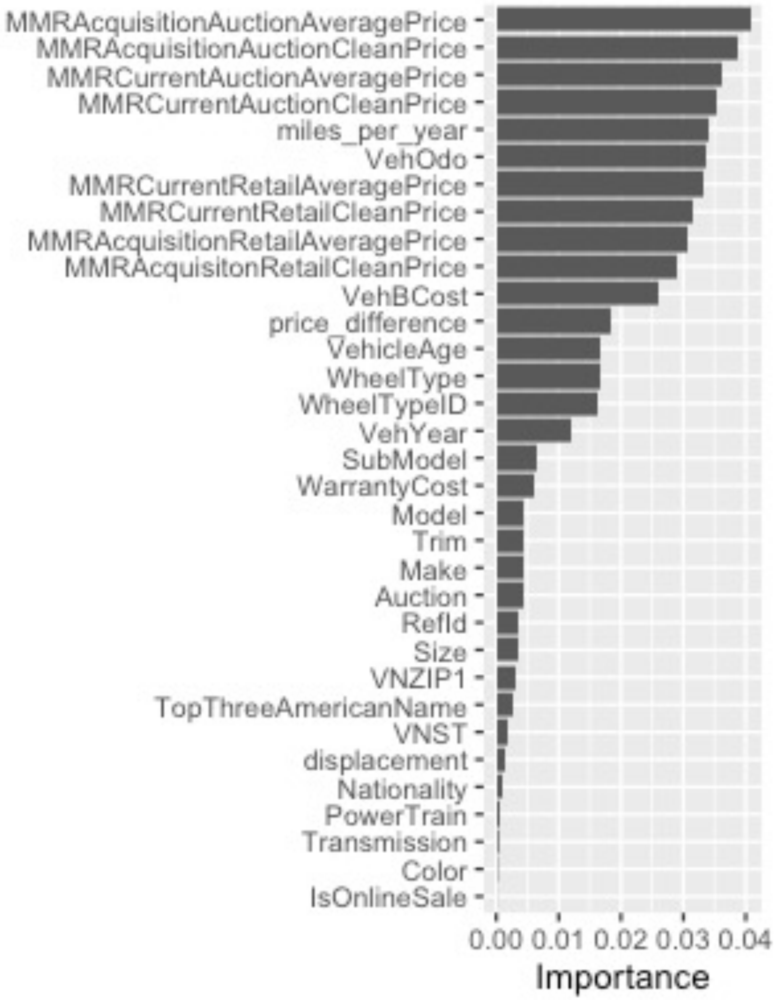
4. Feature Selection

Random Forest 500 Trees Variable Importance

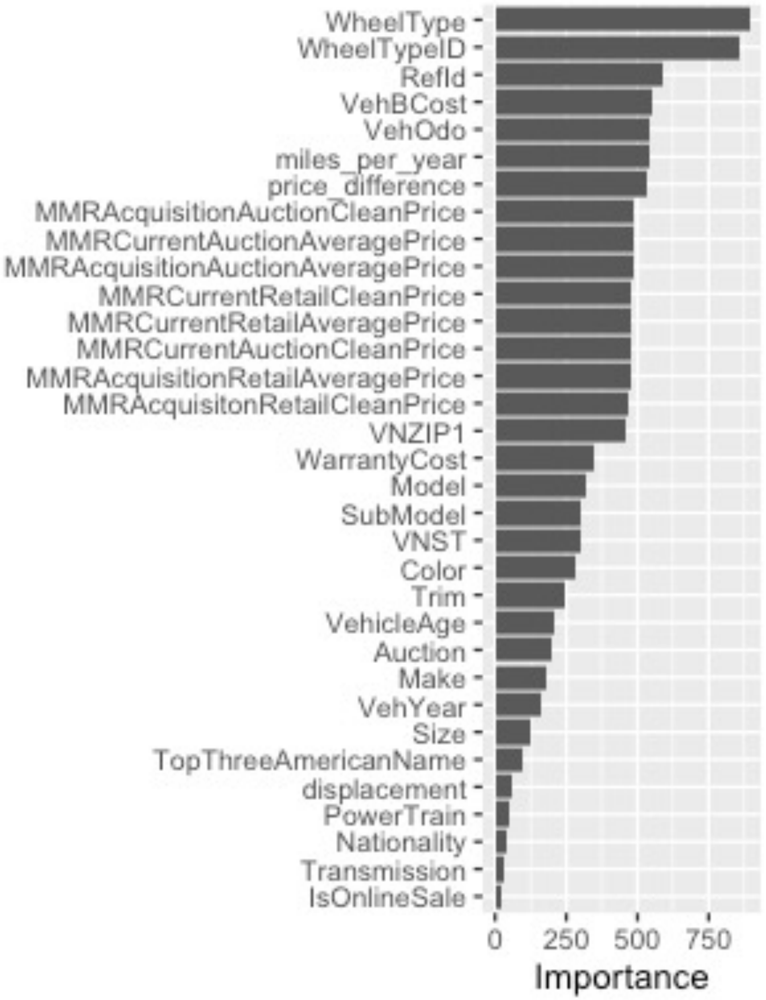
CARVANA

Variable Importance

Permutation Imprtance



Gini Impurity



5. Modelling & Evaluation

Accuracy Sensitivity AUC

5.2 Baseline Performance

Accuracy

0.8870

5.3 Logistic Regression

Features Included:

MMR features, miles_per_year, VehOdo, VehBCost, VehOdo, WheelType, WheelTypeID, and VNZIP1

Model Performance Metrics for Logistic Regression Model

Accuracy	Sensitivity	AUC
.8968	.2458	.7398

5.4 Random Forest

Package ‘ranger’

July 14, 2021

Type Package

Title A Fast Implementation of Random Forests

Version 0.13.1

Date 2021-07-14

Author Marvin N. Wright [aut, cre], Stefan Wager [ctb], Philipp Probst [ctb]

Maintainer Marvin N. Wright <cran@wrig.de>

Description A fast implementation of Random Forests, particularly suited for high dimensional data. Ensembles of classification, regression, survival and probability prediction trees are supported. Data from genome-wide association studies can be analyzed efficiently. In addition to data frames, datasets of class 'gwaa.data' (R package 'GenABEL') and 'dgCMatrix' (R package 'Matrix') can be directly analyzed.

Model Performance Metrics for Random Forest Model

Accuracy	Sensitivity	AUC
.9000	.2425	.7242

CARVANA

5.4 Gradient Boosting

500 Trees

Shrinkage: 0.01

Interaction Depth: 5

5-Fold Cross Validation to Reduce Overfitting and aid in Parameter Tuning

Model Performance Metrics for Gradient Boosting Model

Accuracy	Sensitivity	AUC
.9012	.2475	.7417

Comparison of Results

Model Performance Metrics for All Models. *Darker shade represents better performance.*

Model	Accuracy	Sensitivity	AUC
Logistic Regression	.8968	.2458	.7398
Random Forest	.9000	.2425	.7242
Gradient Boosting	.9012	.2475	.7417

6. Kaggle Submission Result

My Public Score:

0.1375

Top 500 Entries

Best Submission on Kaggle:

0.2672

1st Place Entry

7. Areas of Improvement

- Top Kaggle Submissions used Ensemble Methods
- One-Hot Encoding
- Dataset was heavily class-imbalanced
    - Under sampling or Class Weighting

CARVANA

Thank you for listening

BIOS 635: Final Project Report
Junead Khan